PATENT ABSTRACTS OF JAPAN

(11)Publication number : 02-098775
(43)Date of publication of application : 11.04.1990

(51)Int.Cl. G06F 15/21
G06F 15/18
G06F 15/31

(21)Application number : 63-251334
(22)Date of filing : 04.10.1988

(71)Applicant : NEC CORP
(72)Inventor : YAMANISHI KENJI

(54) METHOD AND DEVICE FOR GENERATING DECISION LIST

(57)Abstract:
PURPOSE: To perform the classification and prediction of data outputted frequently from an information source with high accuracy by picking up the data from decision which maximizes an information gain, and classifying the data preferentially from the decision with high identification capacity.
CONSTITUTION: The decision with high information gain for measuring data is added sequentially as the decision of a decision list as a logical product which regulates the decision of the decision list under a state where the number of characters comprising one logical product is fixed. And a stopping rule based on an error classification rate is provided, and the addition of the decision is stopped at a point where it is satisfied, then, the decision list obtained at that time is outputted. In such a way, it is possible to classify and predict unknown data generated from the same information source as that for the measuring data with high accuracy.

LEGAL STATUS

[Date of request for examination]

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's decision of rejection]

[Date of extinction of right]

⑲日本国特許庁(JP)　⑪特許出願公開

⑫公開特許公報(A)　平2-98775

⑤Int.Cl.⁵	識別記号	庁内整理番号
G 06 F 15/21	Z	7165-5B
15/18		6745-5B
15/31	Z	7050-5B

⑬公開　平成2年(1990)4月11日

審査請求　未請求　請求項の数 8　(全24頁)

⑭発明の名称　決定リストの生成方法及び装置

⑳特　願　昭63-251334

㉒出　願　昭63(1988)10月4日

㉑発明者　山西　健司　東京都港区芝5丁目33番1号　日本電気株式会社内

㉑出願人　日本電気株式会社　東京都港区芝5丁目33番1号

㉔代理人　弁理士　内原　晋

明細書

発明の名称　決定リストの生成方法及び装置

特許請求の範囲

1.{0,1}の値をとる複数の属性とクラスの組として与えられる、雑音を伴う複数の観測データから決定リストを発生させる方法において、1つの論理積を構成する文字の数を固定したもとで、決定リストの決定を規定する論理積として観測データに対する情報利得を最大にするようなものを順次選択しながら付加するステップと、予め定めたストッピングルールによって決定の付加を終了したところで得られる決定リストを最終的に求める決定リストとするステップと、を含むことを特徴とする決定リストの生成方法。

2.請求項1の方法によって一度決定リストを生成するステップと、次に該決定リストの末端から決定項を順次取り除くことによって得られる決定リストの系列の要素を1つ1つに対して、その決定リストを用いて記述される観測データの記述量を計算し、該決定リスト系列の中で記述量を最小にする決定リストを最終的に求める決定リストとするステップと、を含むことを特徴とする決定リストの生成方法。

3.{0,1}の値をとる複数の属性とクラスの組として与えられる、雑音を伴う複数の観測データから決定リストを発生させる方法において、1つの論理積を構成する文字の最大数(kとする)を固定したもとで、決定リストの決定を規定する論理積として観測データに対する情報利得を最大にするようなものを順次選択しながら付加するステップと、予め定めたストッピングルールによって決定の付加を終了したところで得られる決定リストを複数のkの値に対して求めるステップと、各kに対して以上の生成過程を繰り返して生成される決定リストを用いて記述される観測データの記述量を計算して比較し、その中で記述量を最小にする決定リストを最終的に求める決定リストとするステップと、を含むことを特徴とする決定リストの生成方法。

4.請求項1の方法によって各固定されたkに対して決定リストを求めるステップと、該決定リストの末端から決定項を順次取り除くことによって得られる決定リストの系列の要素の1つ1つに対して、その決定リストを用いて記述される観測データの記述量を計算し、該決定リスト系列の中で記述量を最小にする決定リストを各kに対して最良の決定リストとして求めるステップと、複数のk値に対する最良の決定リストを求め、その中で記述量を最小とする決定リストを最終的に求める決定リストとするステップと、を含むことを特徴とする決定リストの生成方法。

5.{0,1}の値をとる複数の属性とクラスの組として与えられる、雑音を伴う複数の観測データから決定リストを発生させる装置において、観測データを記憶する手段と、1つの論理積を構成する文字の数を固定したもとで、決定リストの決定を規定する論理積として観測データに対する情報利得を最大にするようなものを順次選択しながら付加し、予め定めたストッピングルールによって決定の付加を終了したところで得られる決定リストを最終的に求める決定リストとして出力する手段と、を含むことを特徴とする決定リストの生成装置。

6.請求項5の装置によって一度決定リストを生成する手段と、次に該決定リストの末端から決定項を順次取り除くことによって得られる決定リストの系列を発生させる手段と、該系列要素の1つ1つに対して、その決定リストを用いて記述される観測データの記述量を計算する手段と、該記述量及び決定リスト系列を記憶する手段と、該決定リスト系列の中で記述量を最小にする決定リストを最終的に求める決定リストとして出力する手段と、を含むことを特徴とする決定リストの生成装置。

7.{0,1}の値をとる複数の属性とクラスの組として与えられる、雑音を伴う複数の観測データから決定リストを発生させる装置において、観測データを記憶する手段と、1つの論理積を構成する文字の最大数(kとする)を固定したもとで、決定リストの決定を規定する論理積として観測データに対する情報利得を最大にするようなものを順次選択しながら付加し、予め定めたストッピングルールによって決定の付加を終了したところで得られる決定リストを複数のkの値に対して求める手段と、複数のkに対して求められた決定リストを記憶する手段と、各kに対して以上の生成過程を繰り返して生成される決定リストを用いて記述される観測データの記述量を計算する手段と、その中で技術量を最小にする決定リストを最終的に求める決定リストとして出力する手段と、を含むことを特徴とする決定リストの生成装置。

8.請求項5の装置により固定されたkに対して決定リストを求める手段と、該決定リストの末端から決定項を順次取り除くことによって得られる決定リストの系列を発生させる手段と、該系列要素の1つ1つに対して、その決定リストを用いて記述される観測データの記述量を計算する手段と、該記述量を記憶する手段と決定リストの系列を記憶する手段と、該決定リスト系列の中で記述量を最小にする決定リストを各kに対して最良の決定リストとして求める手段と、複数のkの値に対する最良の決定リストを記憶する手段と、複数のkの値に対する最良の決定リストの記述量を比較し、最小とする決定リストを最終的に求める決定リストとして出力する手段と、を含むことを特徴とする決定リストの生成装置。

発明の詳細な説明

(産業上の利用分野)

この発明は{0,1}の値をとる複数の属性とクラスの組として与えられる、雑音を伴う複数の観測データから決定リストを発生させる方法に関する。

(従来の技術)

{0,1}の値をとる複数の属性とクラスの組として与えられる観測データから属性とクラスの構造的な関係を生成し、それを表現するための方法として決定リストによる分類規則生成の方法がある。決定リストの概念については、1987年発行の米国の雑誌「マシンラーニング(Machine Learning)」の2巻の中の229-246頁掲載のR.Lリベスト(R.L.Rivest)による論文「ラーニングデンジョンリ

スト(Learning decision list)」に記載されており、2値データを扱う限りにおいては現在知られている分類規則の表現の中では最も表現能力の高いものであることがわかっている。この論文の中では、雑音の伴わない観測データから、全ての観測データにつじつまを合わせるような決定リストを構成する方法が記載されている。

(発明が解決しようとする課題)

前記論文で示されていた決定リストの生成方法は、雑音の伴わない観測データから全ての観測データにつじつまを合わせるような決定リストを構成するための方法であった。しかし、実際に扱う観測データは一般に雑音を伴うので、全ての観測データを説明出来なくても、同じ情報源から発生するデータを出来るだけ正しく分類するような決定リストの方法が必要であるのだが、このような決定リストを発生させる手段は存在していなかった。

本発明の目的は雑音を伴う観測データから、未知データに対する予測誤差が最小になるような決定リストを自動的に発生させる方法を提供することにある。

(課題を解決するための手段)

本発明による決定リストの発生方法の1つは、1つの論理積を構成する文字の数を固定したもとで、決定リストの決定を規定する論理積として、観測データに対する情報利得を最大にするようなものを順次選択しながら付加するステップと、予め定めたストッピングルールによって決定の付加を終了したところで得られる決定リストを最終的に求める決定リストとするステップとを含むことを特徴とする。

あるいは、上記発明に対して次のような変形を施すことも出来る。すなわち、上の方法によって一度決定リストを生成するステップと、次に該決定リストの末端から決定項を順次取り除くことによって得られる決定リストの系列の要素の1つ1つに対して、その決定リストを用いて記述される観測データの記述量を計算し、該決定リスト系列中で記述量を最小にする決定リストを最終的に求める決定リストとするステップと、を含むことを特徴とする決定リストの発生方法である。(上の2つの方法を「発明1」とする。)

また、本発明によるもう1つの決定リストの発生方法は、1つの論理積を構成する文字の最大数(kとする)を固定したもとで、決定リストの決定を規定する論理積として、観測データに対する情報利得を最大にするようなものを順次選択しながら付加するステップと、予め定めたストッピングルールによって決定の付加を終了したところで得られる決定リストを複数のkの値に対して求めるステップと、各kに対して以上の生成過程を繰り返して生成される決定リストを用いて記述される観測データの記述量を計算して比較し、その中で記述量を最小にする決定リストを最終的に求める決定リストとするステップと、を含むことを特徴とする。

あるいは、上記発明に対して次のような変形を施すことも出来る。すなわち、上の方法によって、各固定されたkに対して決定リストを求めるステップと、該決定リストの末端から決定項を順次取り除くことによって得られる決定リストの系列の要素の1つ1つに対して、その決定リストを用いて記述される観測データの記述量を計算し、該決定リスト系列の中で記述量を最小にする決定リストを各kに対して最良の決定リストとして求めるステップと、複数のkの値に対する最良の決定リストを求め、その中で記述量を最小とする決定リストを最終的に求める決定リストとするステップと、を含むことを特徴とする決定リストの生成方法である。(以下、上の2つの方法を「発明2」とする)。

本発明による決定リストの発生装置の1つは、観測データを記憶する手段と、1つの論理積を構成する文字の数を固定したもとで、決定リストの決定を規定する論理積として観測データに対する情報利得を最大にするようなものを順次選択しながら付加し、予め定めたストッピングルールによって決定の付加を終了したところで得られる決定リストを最終的に求める決定リストとして出力する手段とを含むことを特徴とする。

あるいは、上記発明に対して次のような変形を施すことも出来る。すなわち、上の装置によって一度決定リストを生成する手段と、次に該決定リストの末端から決定項を順次取り除くことによって得られる決定リストの系列を発生させる手段と、該系列要素の1つ1つに対して、その決定リストを用いて記述される観測データの記述量を計算する手段と、該記述量を決定リストの系列を記憶する手段と、該決定リスト系列の中で記述量を最小にする決定リストを最終的に求める決定リストとして出力する手段と、を含むことを特徴とする決定リストの生成装置である。(上の2つの装置を「発明3」とする。)

また、本発明による決定リストの発生装置は、観測データを記憶する手段と、1つの論理積を構成する文字の最大数(kとする)を固定したもとで、決定リストの決定を規定する論理積として観測データに対する情報利得を最大にするようなものを順次選択しながら付加し、予め定めたストッピングルールによって決定の付加を終了したところで得られる決定リストを複数のkの値に対して求める手段と、各kに対して求められた決定リストを記憶する手段と、各kに対して以上の生成過程を繰り返して生成される決定リストを用いて記述される観測データの記述量を計算する手段と、該記述量を記憶する手段と、その中で記述量を最小にする決定リストを最終的に求める決定リストとして出力する手段と、を含むことを特徴とする。

あるいは、上記発明に対して次のような変形を施すことも出来る。すなわち、上の装置により固定されたkに対して決定リストを求める手段と、該決定リストの末端から決定項を順次取り除くことによって得られる決定リストの系列を発生させる手段と、該系列要素の1つ1つに対して、その決定リストを用いて記述される観測データの記述量を計算する手段と、該記述量と決定リストの系列を記憶する手段と、該決定リスト系列の中で記述量を最小にする決定リストを各kに対して最良の決定リストとして求める手段と、複数のkの値に対する最良の決定リストを記憶する方法と、複数のkの値に対する最良の決定リストの記述量を比較し、最小とする決定リストを最終的に求める決定リストとして出力する手段と、を含むことを特徴とする決定リストの生成装置である。(上の2つの装置を「発明4」とする。)

(作用)

先ず、決定リストについて説明する。今、{0,1}に値をとる変数の数をnとし、kはn以下の正の整数であるとする。$L_n=\{x_1,\bar{x}_1,\cdots,x_n,\bar{x}_n\}$ ($\bar{x}_i=1-x_i(i=1,\cdots,n)$)とし、$L_n$の元をリテラルとよび、リテラルの論理積をタームとよぶ。k−DL(n)はT^n_kの元$t_j(j=1,\cdots,r)$と{0,1}にとる値$v_j(j=1,\cdots,r)$の組のリストとして以下のように表されるものの集合を表す。ここに、T^n_kはL_nのうち高々異なるk個のリテラルで表されるターム全体の集合を示し(但し、x_iと同時にx_iを含まない)、最後の関数t_rはtrueを返す定数関数である。

$$(t_1,v_1),\cdots,(t_r,v_r) \qquad (1)$$

1つの決定リストは任意の$x\in X_n$(X_nはn桁のブーリアンベクトルの全体の集合を表す)に対して$t_j(x)=1$となる最初のjに対するv_jの値を示す。(t_i,v_i)を第i決定(あるいは単に決定)とよぶ。例えば、次は3−DT(4)の元である。

$$(x_1\bar{x}_2,1),(x_2\bar{x}_3\bar{x}_5,0),(x_3x_4,1),(\text{true},0) \qquad (2)$$

この決定リストによるデータの分類の決定過程は第7図のように表される。

第7図で示したように、決定リストは"if−then−else　if−then−"型の概念分類規則の一般化となっている。n個の属性と1つのクラスで記述された複数の観測データから同じ情報源より発生するデータを出来るだけ正確に予測するような決定リストを構成するには、どのような方法あるいは装置によって決定リストを発生させたらよいかということが問題であり、この発明に答えるのが本発明である。但し、観測データには雑音が含まれているものとする。先ず、この問題に答える1つの方法として、観測データに対する情報利得(エントロピー、誤分類率、Gini指標等の減少分で測られる)の高い決定から順に決定リストの決定として加えていくようにし、誤分類率に基づくストッピ

ングルールを設けて、これが満足されたところで決定の追加を停止し、そこで得られる決定リストを出力するという方法、あるいはさらにそこで得られる決定リストの末端から決定項を順次取り除くことによって得られる決定リストの系列の要素の1つ1つに対して、その決定リストを用いて記述される観測データの記述量(全てを自己完結的に符号化するための符号長)を計算し、該決定リスト系列の中で記述量を最小にする決定リストを最終的に求める決定リストとする方法が考えられる。情報利得を最大にする決定から拾っていく理由は、識別能力が高い決定から先にデータを分類することにより、情報源から頻出するデータについて、より精度の高い分類・予測が行えるからである。また、記述量を最小にする決定リストを選ぶことの理由は、観測データをモデルを用いて自己完結的な符号化を行う際に、全データをより圧縮しうるモデル(この場合は決定リスト)は、同じ情報源より発生する未知の観測データに対し、より正確な予測を行えるモデルであるということが漸近的に成立するからである。この理論的裏付けに関しては、特に統計モデルに適用する場合に関しては1986年発行の米国の雑誌アナルスオブスタティスティクス(Annals of Statistics)の14巻の1080-1100頁掲載のJ.リサネン(J.Rissanen)による論文「ストキャスティックコンプレキシティーアンドモデリング(Stochastic complexty and modeling)」にMDL基準として述べられている。しかし、本発明では、分類規則としての決定リストといった論理型概念学習の最適モデル化にMDL基準を用いているところに新規性をおいている。以上が発明1の方法の原理であり、同じ原理を装置の形で実現するのが発明3である。上述の情報利得を以下に正確に定義する。先ず、以下のように記号を定める。

[記号] 以下、対数の底は全て2とする。

・k,n∈N(自然数全体)は固定とする。

・S≜ {<i,x,c>(学習データ): i(対象番号)∈N, x(属性値)∈{0,1}^{n},c(クラス)∈{0,1}}: 学習データの集合

(属性値は属性変数$x_1,\cdots,X_n$に対するデータの値を示す。)

・$SA(t)=\{<i,x,c>\in S:t(x)=1\},t\in T^n_k$

$SB(t)=\{<i,x,c>\in S:t(x)=0\},t\in T^n_k$

・$A(t)=\#SA(t)$(以下,#Wは集合Wの元の総数を表す)

$B(t)=\#SB(t)$

・$A^+(t)=\#\{<i,x,c>\in SA(t): v=1\},t\in T^n_k$

$A^-(t)=\#\{<i,x,c>\in SA(t): v=0\},t\in T^n_k$

$B^+(t)=\#\{<i,x,c>\in SB(t): v=1\},t\in T^n_k$

$B^-(t)=\#\{<i,x,c>\in SB(t): v=0\},t\in T^n_k$

ここに、$A(t)=A^+(t)+A^-(t)$,$B(t)=B^+(t)+B^-(t)$とする。

$$I(t)\triangleq\frac{A(t)}{A(t)+B(t)}I_A(t)+\frac{B(t)}{A(t)+B(t)}I_B(t) \quad (3)$$

・$\Delta I(t)\triangleq I_D(t'')-I(t)$ (4)

を決定(t,v)による情報利得とよぶ。

(tは決定リストにおいてtの直前に現れる決定のタームとする。)

ここで、$I_A(t)$,$I_B(t)$の選び方としては次のようなものを考えることが出来る。

(i) エントロピー

$$I_A(t)\triangleq-\frac{A^+(t)}{A(t)}\log\frac{A^+(t)}{A(t)}-\frac{A^-(t)}{A(t)}\log\frac{A^-(t)}{A(t)}$$

$$I_B(t)\triangleq-\frac{B^+(t)}{B(t)}\log\frac{B^+(t)}{B(t)}-\frac{B^-(t)}{B(t)}\log\frac{B^-(t)}{B(t)} \quad (5)$$

(ii) Gini指標

$$\cdot I_A(t)\triangleq\frac{A^+(t)}{A(t)}\cdot\frac{A^-(t)}{A(t)}$$

$$\cdot I_B(t)\triangleq\frac{B^+(t)}{B(t)}\cdot\frac{B^-(t)}{B(t)} \quad (6)$$

(iii) 誤分類率

$$\cdot I_A(t)\triangleq\frac{\min\{A^+(t),A^-(t)\}}{A(t)}$$

$$\cdot I_B(t)\triangleq\frac{\min\{B^+(t),B^-(t)\}}{B(t)} \quad (7)$$

次に、決定リストを用いた全観測データの記述量を正確に定義する。データの記述量は、決定リスト自体の記述量と決定リストによって分類されるデータの中の例外の記述量との和として与え

られる。例えば、(2)の決定リストつにいては、$\#T^5_3=131$である(定数関数trueに対応するタームを0としてこれも数える)から、(2)の最初の決定のタームは1/131の確率で選ばれているので、これを自己完結的に符号化するためには、log(131)bitsの符号長が必要である。この場合、符号化方法としては、Huffman符号化を用いる。また、1つの決定に対しては、タームを真にするようなデータに1を割るけるかを0を割り付けるかも記述しなければならず、これには1bit必要であるから、結局(2)の最初の決定に関する記述量として、log(131)+1bits必要である。次の決定に関しては、T^5_3の最初の決定に用いられたタームを除く130個のタームの中から決定に必要なタームが選ばれるのであるから、同様にして、log(130)+1bitsの記述量が必要である。同様にして、各決定に対する記述量を求め、それらの総和を計算することにより決定リスト自体の記述量か計算できる。(2)に対しては全部で(log(131)+1)+(log(130)+1)+(log(129)+1)+(log(128)+1)bits必要である。また、例外データの記述量は、例えば、1つの決定に対して、その決定値が1であるとして、その決定におけるタームを真にするデータが7つであり、そのうちクラスが1であるものの数が5,0であるものの数が2であるとして、対象番号の若い順にデータのクラスを記述すると、1101101であったとする。このとき、この系列を自己完結的に符号化するのに必要な符号長は、$\log(7+1)+\log({}_7C_2)$bitsまたは初めから例外は必ず$\lfloor(7+1)/2\rfloor$個以下であることが分かっていれば、$\log(\lfloor(7+1)/2\rfloor+1)+\log({}_7C_2)$bitsである。一般に、系列の長さを$N,b=b_1$または$b_2$、ここに$b_1=N,b_2=\lceil(N+1)/2\rceil$とし、例外の数をhすることにより、例外記述に必要な記述量は

$$\log(b+1)+\log({}_NC_h)\text{bits} \tag{8}$$

または、

$$L_N(h)+\log({}_NC_h)\text{bits} \tag{9}$$

で与えられる。ここに、$L_N(h)$は$\{0,1,\cdots,N\}$に含まれる自然数の自己完結的な符号化を行うときの符号長を表し、次に満たす。

$$L_N(0)=1$$

$$L_N(k)=1+\log k+\log\log k+\cdots+C_N(k>1) \tag{10}$$

但し、上の和は正の項のみに対してとられるものであり、C_Mは$\Sigma_{k=0}{}^M 2^{-L_N(k)}=1$を満たす実数である。

以上に与えた方法によって算出される決定リスト自身の記述量と例外データの記述量との和が決定リストを用いることによる全学習データの記述量である。上で述べた情報利得と記述量の概念を用いて、所与の観測データに対する決定リストの最適化を行うことが出来る。例えば、属性数が6の次のような学習データを考える。

(以下余白)

対象	x_1	x_2	x_3	x_4	x_5	x_6	クラス
1	0	1	1	0	1	1	1
2	1	1	1	0	0	0	1
3	0	1	0	1	1	0	1
4	0	1	1	0	0	1	0
5	1	1	1	0	1	1	1
6	1	0	0	0	0	1	0
7	0	1	1	1	1	0	1
8	1	0	1	0	1	1	1
9	0	1	0	1	1	0	1
10	1	1	1	0	1	1	1
11	0	1	1	0	0	1	1
12	0	1	1	1	0	0	1
13	1	0	0	0	0	0	1
14	0	1	1	1	1	1	0
15	0	1	1	1	1	0	0
16	1	1	1	0	0	1	1
17	0	1	1	0	1	1	0
18	0	1	0	1	0	1	1
19	1	0	1	0	0	0	0
20	0	1	1	1	0	0	0
21	0	1	1	0	0	0	1
22	0	1	0	0	0	0	1
23	1	1	1	1	1	1	0
24	1	0	1	0	0	0	1
25	0	1	0	1	1	0	1
26	1	0	1	1	0	0	1
27	0	1	1	0	0	1	1
28	1	0	0	0	0	1	1
29	0	1	1	0	0	0	0
30	0	1	0	1	1	0	1
31	0	0	0	0	0	1	0
32	1	1	1	1	1	0	0

これらの中に、3と9、4と11、12と20、19と24のデータに矛盾を含んでいる。発明1によれば、先ず、kの値を固定したときの決定リストを、情報利得を最大にする決定から付加して行く方法で構成し、予め定めたストッピングルールで停止して、そこで得られる決定リストを求める決定リストとすることにより、次のような決定リストが得られる(k=4,5に対し、それぞれDL*(4),DL*(5)とかく)。但し、情報利得はエントロピーを用いるものとし、このときの決定付加の停止条件として、前決定における決定値をv、情報利得を最大にするタームをtとして、次のi),ii),iii)の条件を採用するものとする。

i) $\min\{A^+(t),A^-(t)\}/A(t)>\alpha(=0.25)$ならば、

$B^+(t'')<B^-(t'')$かつv=1あるいは$B^+(t'')>B^-(t'')$かつv=0(true,1−v)を決定として右に付け加えて出力し、停止する。

ii) $\min\{B^+(t),B^-(t)\}/B(t)\leq\beta(=0.29)$かつ、

$B^+(t)<B^-(t)$かつv=1あるいは$B^+(t)>B^-(t)$かつv=0ならば、(true,1−v)を決定として右に付け加えて出力し、停止する。

iii) まだ決定されていない観測データがもうなければ、

(true,1−v)を決定として右に加えて出力し、停止する。

DL*(4)は次の通り

$(x_2x_3x_5\bar{x}_6,1)(x_1\bar{x}_2\bar{x}_4,0)(\bar{x}_1x_2x_4\bar{x}_6,1)$

$\cdot(x_2x_3\bar{x}_4\bar{x}_5,1)$(true,0)

DL*(5)は次の通り

$(x_2x_3x_5x_6,1)(x_1\bar{x}_2\bar{x}_4,0)(\bar{x}_1x_2x_4x_5\bar{x}_6,1)$

$\cdot(\bar{x}_1x_2x_3\bar{x}_4\bar{x}_5,1)(x_1x_3\bar{x}_5\bar{x}_6,1)$(true,0)

また、発明3によれば、観測データを記憶する手段と、情報利得を最大にする決定から順に付加して、予め定めているストッピングルールの条件が満たされたときに得られる決定リストを求める最終的な決定リストとして出力する手段を具備している装置によって、k=4,5のときにはそれぞれDL*(4),DL*(5)が発生させられる。

尚、以上の決定リストの発生方法並びに装置では、固定されたkに対してのみk−DL(n)が発生させることが出来る。一般にkの値が大きければ大きいほど表現能力が高くなり、多次元のデータ空間の分割の精度も細かくなるが、観測データには一般に雑音が入っているので、kが必要以上に大きいと、統計的な揺らぎに過敏な決定リストを構成してしまうことになり、その場合、未知データに対する分類予測誤差は返って大きくなる。従って、同じ情報源から発生する未知データに対する予測誤差を最小にするような最適なkの値が存在するはずであり、このようなkに対する決定リストを発生させる方法並びに装置が必要となる。上述の意味で最適な決定リストを発生させるためには、様々なkの値に対して、発明1の方法または発明3の方法によって決定リストを発生させてから、それらを用いて記述される観測データの記述量を計算して比較し、記述量を最小にするような決定リストを最終的に求める決定リストとして求める方法並びに装置が考えられる。この理論的根拠も前出のJ.リサネンによる論文に示されているMDL基準の考え方に依るものである。以上が発明2の方法の原理であり、同じ原理を装置の形で実現するのが発明4である。例えば、前出の例では、$\#T^6_4=473$, $\#T^6_5=665$であり、各決定における(決定される対象の数、例外の数)は、

DL*(4)で、

(6,0),(6,1),(8,2),(7,2),(5,1)

DL*(5)で、

(6,0),(6,1),(6,1),(5,1),(2,0),(7,1)

であるから、記述量は、例外の記述量の計算方法としては$b=b_1$として(7)を用いることにすれば、次のように求められる。

DL*(4)で、

モデルの記述量 = 48.398 bits

例外の記述量 = 24.750 bits

L(4) = 73.148 bits

DL*(5)で、

モデルの記述量 = 55.230 bits

例外の記述量 = 20.621 bits

L(5) = 75.851 bits

従って、MDL基準の下ではDL*(4)がDL*(5)よりも適当なモデルであると言うことが出来る。発明2によると、k=4とk=5に対して発明1の方法で1度決定リストを発生させ、さらに各kに対する記述量を計算し、それらを比較して小さい記述量を与える決定リストとして、DL*(4)を発生させることが出来る。発明4によると、k=4とk=5に対して発明3の装置で1度決定リストを発生させる手段と、各kに対する記述量を計算する手段と、それら及び各kに対する決定リストを記憶する手段と、記述量を比較し、最小にする決定リストを出力する手段を具備していれば、DL*(4)を発生させることが出来る。

(実施例)

次に、本発明について図面を参照して詳細に説明する。以下、記号は(作用)の項に従う。

第1図は発明1の実施例を説明するフローチャートである。startでは、対象番号と2値のn個の属性と2値のクラスからなる観測データの全集合を初期値とする集合Sと、固定されたn,kに対してT^n_kを初期値とする集合Tと、初期ターム$t''=0$(ターム0はT^n_kの中に含まれないが、全データを全て偽にするタームとして定める。また、このタームを用いた決定は出力時点では省略されるものとする)と、初期決定値v=0が与えられている。ステップ11で、ストッピングルールの条件として、最初から与えられている1以下の正の実数α,βに対して、

i) S=空集合

ii) $\min\{A^+(t''),A^-(t'')\}/A(t'')>\alpha$

$B^+(t'')>B^-(t'')$かつv=1あるいは$B^+(t'')<B^-(t'')$かつv=0

iii) $\min\{B^+(t''),B^-(t'')\}/B(t'')\leq\beta$かつ

$B^+(t'')>B^-(t'')$かつv=1あるいは$B^+(t'')<B^-(t'')$かつv=0

を設けて、i) ii) iii) の順にこれらの条件の1つでも当てはまるかどうかを調べる。これらの条件のうち1つでもあてはまるものがあれば、ステップ12に進み、最終決定として(true,1−v)を右に付加して、それまで付加してきた決定と併せて出力して終了する。これらの条件がいずれも満たされてないならば、ステップ13に進み、(4)で定められている情報利得をタームtの関数とみなしたときに、これを最大にするタームtをTの中から決定する。(これをt^*とする。)情報利得最大のタームが複数ある場合は、その中でランダムにt^*を選ぶ。次に、ステップ14で、$A^+(t^*)>A^-(t^*)$ならば、v=1とし、$A^+(t^*)<A^-(t^*)$ならば、v=0とし、(t^*,v)を決定リストの決定として右に付け加える。次に、ステップ15で、タームtを真にするデータをSから除いた集合を改めてSとし、タームtをTから除いた集合を改めてTとし、t^*を改めてt''とし、ステップ11に戻る。

第2図は発明1のもう1つの実施例を説明するフローチャートである。ステップ21から25まではそれぞれ、第1図のステップ11から15までと同じ機能を果たし、ステップ22の出力として、第1図の出力と同じ決定リストが得られる。第2図では、ステップ22の出力としての決定リスト($DL_{max}(k)$とする)に対して、さらにステップ26で、$DL_{max}(k)$を用いて記述できる観測データの記述量を(作用)の項で述べたような方法で決定リスト自身の記述量と決定に対する例外データの記述量との和として計算し、これをL_0とする。ここで、記述量は、決定リスト自身の記述量と決定に対する例外データの記述量との単純な和ではなく、重み付き和、すなわち、λ(決定リストの記述量)+(1−λ)(例外データの記述量)(0<λ<1)として計算する場合もあるとする。以下の記述量計算においても同様である。次に、ステップ27で、$DL_{max}(k)$の末端(最終)決定から刈り込みを行う。具体的には、$V(1)=DL_{max}(k)$として、j≧1に対し、V(j)の右から2つめの決定を(t,v),(true,1−v)とするとき、これらを(true,1−v^*)に置き換えたものをV(j+1)とする。但し、t'をtの直前の決定に用いられたタームとすると、v^*は次のように与えられる。

$$v^*=\begin{cases}1 & A^+(t')\geq A^-(t')\text{の場合}\\ 0 & A^+(t')<A^-(t')\text{の場合}\end{cases}$$

次に、ステップ28で、V(j+1)を用いて記述される観測データの記述量を(作用)の項に示した方法で

計算し、これをL_{j+1}とする。次に、ステップ29に進み、V(j+1)からさらに刈り取る決定が残っているかどうかを判断し、残っていなければ、ステップ31に進み、記述量全てのL_jの値を比較して最小になるものに対するV(j)を出力し、終了する。ここで、もし、最小値が複数存在すれば、その最小値に対応するV(j)を複数出力し、終了する。また、ステップ29においてV(j+1)から刈り取る決定が残っていれば、ステップ30に進み、jをj+1としてステップ27に戻る。

第3図は発明2の実施例を説明するフローチャートである。startにおいては、対象番号と2値のn個の属性と2値のクラスとからなる観測データの全集合を初期値とする集合Sと、固定されたn、と複数のkに対して(kは1つの決定に用いられる論理積を構成する文字の最大数であり、kのとりうる数の総数をMとする)、T^n_kを初期値とする集合T(k)と、初期タームt"=0と、初期決定値v=0が与えられている。

先ず、各kに対して、発明1の方法で決定リストを発生させる。$k_1<\cdots<k_M$を対象にするkとして、k_jに対応する決定リスト発生のステップを31とする。ステップ31は第1図の全体或は第2図の全体である。次に、発生した決定リストに対して、その決定リストを用いて記述されるデータの記述量を計算する。第2図で示された方法をステップ31にする場合はこのステップは省略される。k_iに対応する記述量計算のステップをステップ32で表す。次に、ステップ33各kの値に対して計算された記述量を比較して、最小値を求める。ここで、記述量は、決定リスト自身の記述量と決定に対する例外データの記述量との単純な和ではなく、重み付き和、すなわち、λ(決定リストの記述量)+(1−λ)(例外データの記述量)(0<λ<1)として計算する場合もあるとする。次に、ステップ34で、これらの中で最小なkの値に対応する決定リストを出力し、終了する。第4図は発明3の装置を示すブロック図である。対象番号と2値のn個の属性と2個のクラスとからなる観測データの全集合を入力として、これを41の記憶装置で一旦記憶する。制御信号発生装置から発生する制御信号の指令によって記憶装置から観測データの一部が情報利得最大ターム決定回路42に送られる。最初に供給されるデータは観測データ全部である。情報利得最大ターム決定回路42は初期状態のパラメータとして、$T=T^n_k$(kは固定)、t"=0を有しており、

i) S＝空集合

ii) $\min\{A^+(t"),A^-(t")\}/A(t")>\alpha$

$B^+(t")>B^-(t")$かつv＝0あるいは$B^+(t")<B^-(t")$かつv＝1

iii) $\min\{B^+(t"),B^-(t")\}/B(t")\leq\beta$かつ、

$B^+(t")>B^-(t")$かつv＝0あるいは$B^+(t")<B^-(t")$かつv＝1

の条件をi) ii) iii) の順に調べ、1つでも当てはまるような場合は、信号P及びt"を42に縦属する決定付加回路43に送る。また、上の条件がいずれも満たされない場合には、固定されたkの値に対して、Tの中のタームについて、記憶装置から供給されるデータに対して(4)で与えられる情報利得を計算し、これを最大にするようなタームt*を決定し、状態をt"=t*,T=T−t*に変えて、t*を決定付加回路43に送る。決定付加回路43は、情報利得最大ターム決定回路42の出力と記憶装置から供給されるデータ(42に供給されるデータと同じ)を入力として、42からt*が送られてきた場合は$A^+(t^*),A^-(t^*)$を計算して、$A^+(t^*)>A^-(t^*)$ならば、v＝1,$A^+(t^*)<A^-(t^*)$ならば、v＝0とし、(t*,v)を決定リストの決定として右に付け加え、次に記憶装置41がS−SA(t*)を42に供給するように指令する制御信号を発生させる信号を、制御信号発生装置44に送る。44は記憶装置がS−SA(t*)を42に供給するように指令する制御信号を発生させる信号を41に送る。41は44からの指令を受けて、S−SA(t*)を41に供給する。44の指令を受けて新たに記憶装置41からデータが供給されてきた。42は同じ動作を繰り返す。43に42から信号Pが送られてきた場合には、$A^+(t^*)>A^-(t^*)$ならば、v＝1,$A^+(t^*)<A^-(t^*)$ならば、v＝0として、(true,1−v)を最後の決定として決定リストの右に付

け加えて、回路全体の出力としてこれまで決定を付加して得られる決定リストを出力し、回路全体の動作を終了する。

第5図は発明3の決定リストの発生装置のもう1つの実施例を示すブロック図である。回路51,52,54はそれぞれ、第4図の回路41,42,44と同じ入出力機能を果たす。しかし、回路53の出力としては、第1図の出力と同じ決定リストにくわえて、各決定に用いられたタームのそれぞれに対するA⁺(t),A⁻(t)の値も出力される。刈り込み及び記述量計算回路55は、回路53の出力を入力として、それを末端の決定から順に刈り込んで、それらのそれを用いてデータを記述するときの記述量を合わせて計算する。ここで、記述量は、決定リスト自身の記述量と決定に対する例外データの記述量との単純な和ではなく、重み付き和、すなわち、λ(決定リストの記述量)+(1−λ)(例外データの記述量)(0<λ<1)として計算する場合もあるとする。具体的には、右から2つの決定を(t,v),(true,1−v)とするとき、これらを(true,1−v*)に置き換えるといった操作を刈り取る決定が無くなるまで繰り返す。但し、t'をtの直前の決定に用いられたタームとすると、v*は次のように与えられる。

$$v^*=\begin{cases}1 & A^+(t')\geqq A^-(t')\text{の場合}\\ 0 & A^+(t')<A^-(t')\text{の場合}\end{cases}$$

55はそれぞれ刈り込みによって得られた決定リストのそれぞれと、それぞれに対して計算された記述量を合わせて、第2記憶装置56に送る。第2記憶装置は55から送られてきた入力を記憶する。また、55は刈り込まれたそれぞれの決定リストの記述量だけを記述量比較回路57にも送る。57は刈り込まれた決定リスト達の記述量を比較し、最小値を求め、第2記憶装置56から記述量を最小にする決定リストを出力させることを指示する制御信号を送ることを指令する信号を、第2制御信号発生回路58に送り込む。58は57の指令を受けて、第2記憶装置56から記述量を最小にする決定リストを出力させることを指示する制御信号を第2記憶装置56に送る。第2記憶装置56は58の指令を受けて、記述量を最小にする決定リストを出力する。

第6図は発明4の決定リストの発生装置の実施例を示すブロック図である。観測データを装置全体の入力として、該入力は先ず、DL*(k)発生回路61に送り込まれる。61は固定されたkに対して、発明3の装置と同じ回路によって決定リストを発生させ(固定されたkに対して発生された決定リストをDL*(k)とする)、これを第1制御信号発生装置62より発生する制御信号の指示に従って複数のkの値に対して実行する。61は各kに対するDL*(k)と各決定に用いられたタームのそれぞれに対するA⁺(t),A⁻(t)の値を順に記述量計算回路63に送り込み、各kに対するDL*(k)を記憶装置64に送り込む。63は61の出力を入力として、それぞれの記述量を計算し、その記述量を記述量比較回路65及び記憶装置64に送り込む。ここで、記述量は、決定リスト自身の記述量と決定に対する例外データの記述量との単純な和ではなく、重み付き和、すなわち、λ(決定リストの記述量)+(1−λ)(例外データの記述量)(0<λ<1)として計算する場合もあるとする。記憶装置64は61及び63の出力を入力として、これらを記憶する。65は63の出力を入力として、65は各kに対するDL*(k)の記述量を比較し、最小値を求め、記憶装置64から記述量を最小にする決定リストを出力させることを指示する制御信号を送ることを指令する信号を、第2制御信号発生回路66に送り込む。66は65の指令を受けて、記憶装置64から記述量を最小にする決定リストを出力させることを指示する制御信号を記憶装置64に送る。記憶装置64は66の指令を受けて、記述量を最小にする決定リストを出力する。

(発明の効果)

発明1及び発明3によれば、固定されたkの値(決定を規定する論理積に現れる文字の最大数)に対して、観測データと同じ情報源から発生する未知データを出来るだけ正確に分類することが理論的に保証された決定リストが発生できる。発明2及び発明4によれば、さらに、kの値を動かして得られる得られる広い決定リストの集合の中から、観測データと同じ情報源から発生する未知データを出

来るだけ正確に分類することが理論的に発生された決定リストを選ぶことが出来る。

図面の簡単な説明

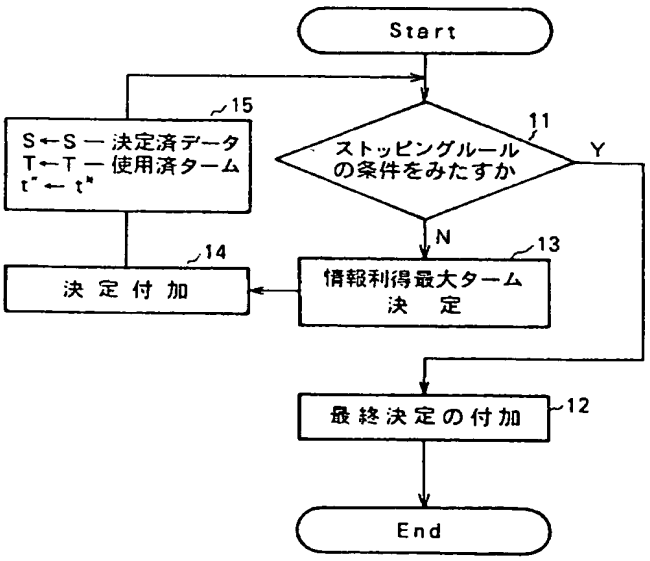
第1図は発明1の決定リストの生成方法を示すフローチャート、第2図は発明1の決定リストの生成方法で、第1図に示した方法の変形版のフローチャート、第3図は発明2の決定リストの生成方法を示すフローチャート、第4図は発明3の決定リストの生成装置を示すブロック図、第5図は発明3の決定リストの生成装置で、第4図に示した方法の変形版のブロック図、第6図は発明4の決定リストの生成装置を示すブロック図、第5図は発明4の決定リストの生成装置で、第6図に示した方法の変形版のブロック図、第7図、第8図は本発明の作用を説明するための図である。

図において、

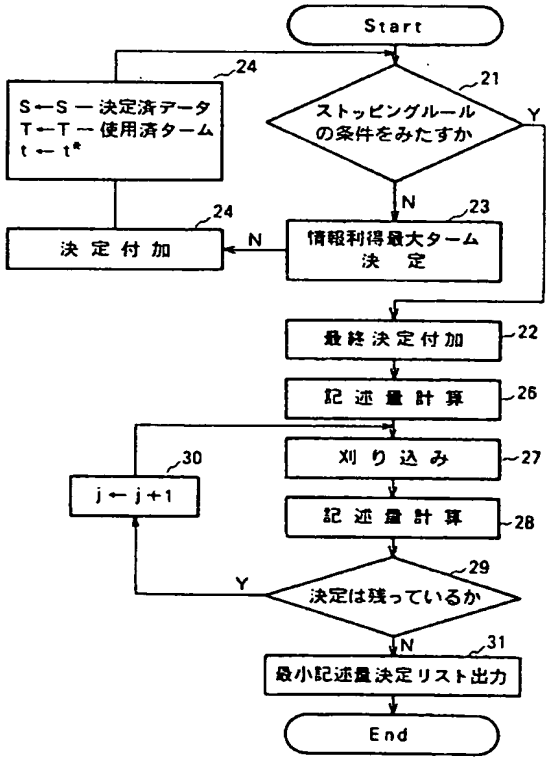
41: 記憶装置、42: 情報利得最大ターム決定回路、43: 決定付加回路、44: 制御信号発生装置、51: 第1記憶回路、52: 情報利得最大ターム決定回路、53: 決定付加回路、54: 第1制御信号発生装置、55: 刈り込み及び記述量計算回路、56: 第2記憶回路、57: 記述量比較回路、58: 第2制御信号発生装置、61: DL*(k)発生回路、62: 第1制御信号発生装置、63: 記述量計算回路、64: 記憶装置、65: 記述量比較回路、66: 第2制御信号発生装置。

代理人 弁理士 内原 晋

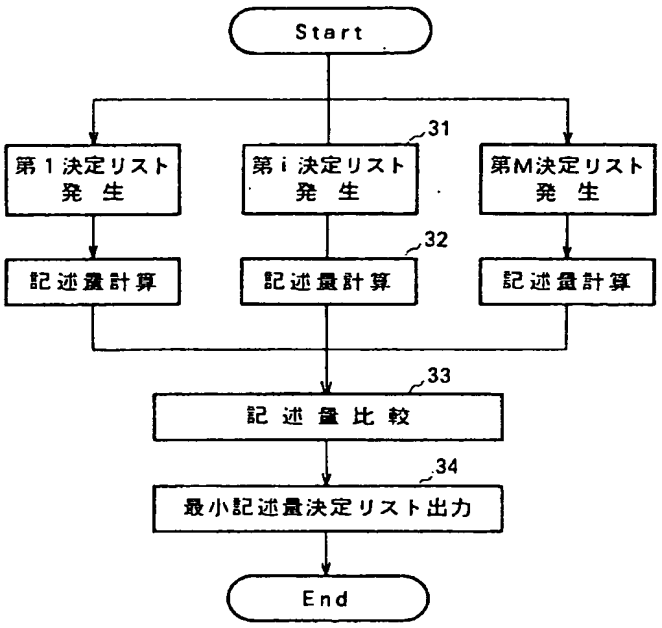
第 1 図



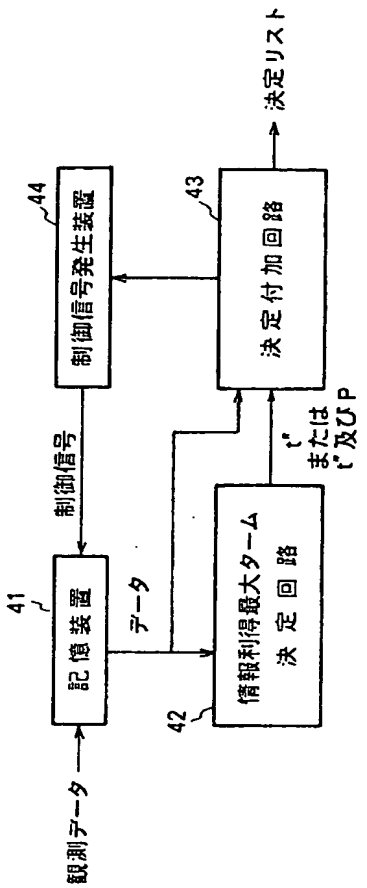
第 2 図



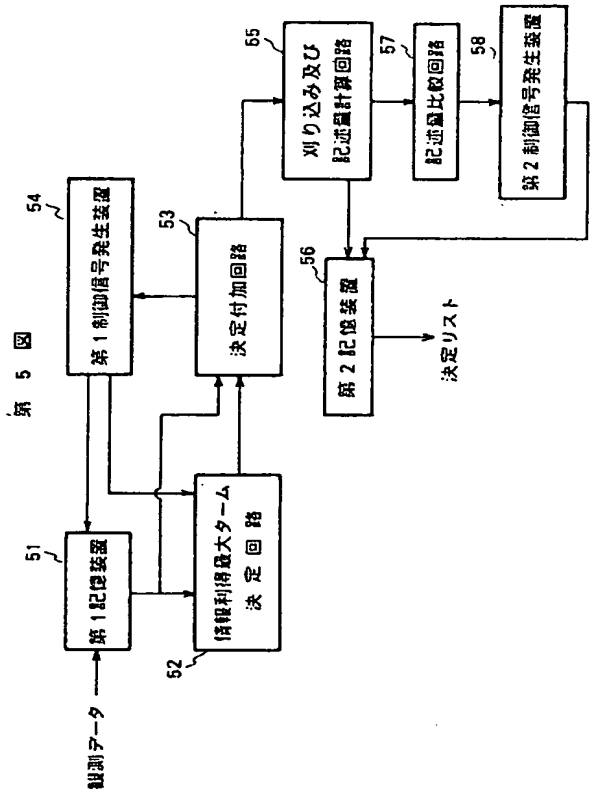
第 3 図



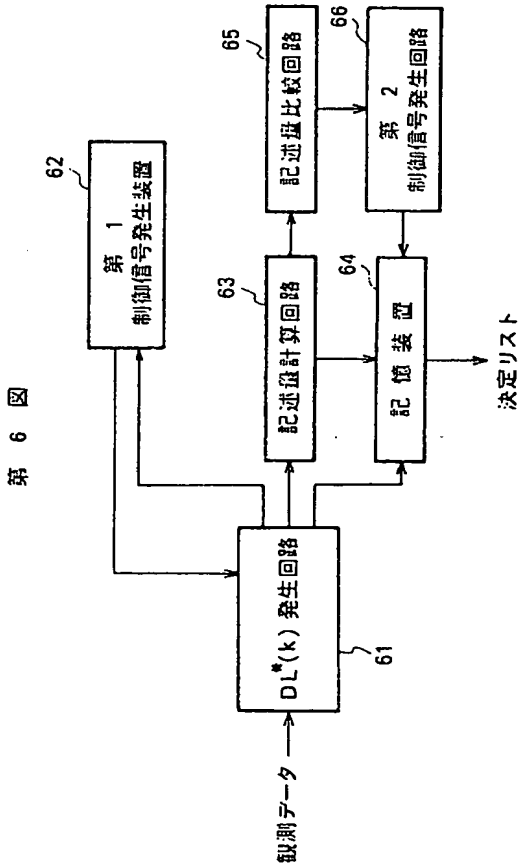
第 4 図



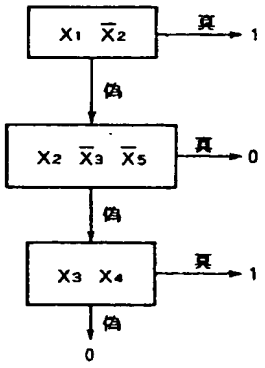
第 5 図



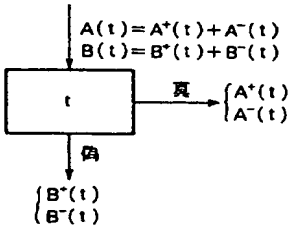
第 6 図



第 7 図



第 8 図



5.補正の対象

明細書の特許請求の範囲の欄

明細書の発明の詳細な説明の欄

明細書の図面の簡単な説明の欄

図面

6.補正の内容

(1)明細書全文を別紙のとおり補正する。(補正の対象に記載した事項以外内容に変更なし)

(2)本願添付図面第2図、第6図を別紙図面のとおり補正する。

(3)本願添付図面として、別紙の第9図(a),(b)を追加する。

代理人 弁理士 内原 晋

手 続 補 正 書 (自発)

1.11.-8

平成 年 月 日

特許庁長官 殿

1. 事件の表示 昭和 63年 特許願 第 251334号

2. 発明の名称

決定リストの生成方法及び装置

3. 補正をする者

事件との関係 出 願 人

東京都港区芝五丁目33番1号

(423) 日本電気株式会社

代表者 関 本 忠 弘

4. 代 理 人 方式 審査 (関)

〒108 東京都港区芝五丁目37番8号 住友三田ビル

日本電気株式会社内 弁理士 内原晋

(6591) 弁理士 内 原

電話 東京 (03) 456−3111 (大代表)

(連絡先 日本電気株式会社 特許部)

1.11.10

別紙

明 細 書

発明の名称 決定リストの生成方法及び装置

特許請求の範囲

1.複数の属性とクラスの組として与えられる、複数の観測データから決定リストを発生させる方法において、1つの論理積を構成する文字の数を固定したもとで、決定リストの決定を規定する論理積として観測データに対する情報利得を最大にするようなもの順次選択しながら付加するステップと、予め定めたストッピングルールによって決定の付加を終了したところで得られる決定リストを最終的に求める決定リストとして出力するステップと、を含むことを特徴とする決定リストの生成方法。

2.請求項1の方法によって一度決定リストを生成するステップと、次に該決定リストの末端から決定項を順次取り除くことによって得られる決定リ

ストの系列の要素を1つ1つに対して、その決定リストを用いて記述される観測データの記述量を計算し、該決定リスト系列の中で記述量を最小にする決定リストを最終的に求める決定リストとして出力するステップと、を含むことを特徴とする決定リストの生成方法。

3.複数の属性とクラスの組として与えられる、複数の観測データから決定リストを発生させる方法において、1つの論理積を構成する文字の最大数(kとする)を固定したもとで、決定リストの決定を規定する論理積として観測データに対する情報利得を最大にするようなものを順次選択しながら付加するステップと、予め定めたストッピングルールによって決定の付加を終了したところで得られる決定リストを複数のkの値に対して求めるステップと、各kに対して以上の生成過程を繰り返して生成される決定リストを用いて記述される観測データの記述量を計算して比較し、その中で記述量を最小にする決定リストを最終的に求める決定リストとして出力するステップと、を含むことを特徴とする決定リストの生成方法。

4.請求項1の方法によって各固定されたkに対して決定リストを求めるステップと、該決定リストの末端から決定項を順次取り除くことによって得られる決定リストの系列の要素1つ1つの対して、その決定リストを用いて記述される観測データの記述量を計算し、該決定リスト系列の中で記述量を最小にする決定リストを各kに対して最良の決定リストとして出力するステップと、複数のk値に対する最良の決定リストを求め、その中で記述量を最小とする決定リストを最終的に求める決定リストとするステップと、を含むことを特徴とする決定リストの生成方法。

5.複数の属性とクラスの組として与えられる、複数の観測データから決定リストを発生させる装置において、観測データを記憶する手段と、1つの論理積を構成する文字の数を固定したもとで、決定リストの決定を規定する論理積として観測データに対する情報利得を最大にするようなものを順次選択しながら付加し、予め定めたストッピングルールによって決定の付加を終了したところで得られる決定リストを最終的に求める決定リストとして出力する手段と、を含むことを特徴とする決定リストの生成装置。

6.請求項5の装置によって一度決定リストを生成する手段と、次に該決定リストの末端から決定項を順次取り除くことによって得られる決定リストの系列を発生される手段と、該系列要素の1つ1つに対して、その決定リストを用いて記述される観測データの記述量を計算する手段と、該記述量及び決定リスト系列を記憶する手段と、該決定リスト系列の中で記述量を最小にする決定リストを最終的に求める決定リストとして出力する手段と、を含むことを特徴とする決定リストの生成装置。

7.複数の属性とクラスの組として与えられる、複数の観測データから決定リストを発生させる装置において、観測データを記憶する手段と、1つの論理積を構成する文字の最大数(kとする)を固定したもとで、決定リストの決定を規定する論理積として観測データに対する情報利得を最大にするようなものを順次選択しながら付加し、予め定めたストッピングルールによって決定の付加を終了したところで得られる決定リストを複数のkの値に対して求める手段と、複数のkに対して求められた決定リストを記憶する手段と、各kに対して以上の生成過程を繰り返して生成される決定リストを用いて記述される観測データの記述量を計算する手段と、その中で記述量を最小にする決定リストを最終的に求める決定リストとして出力する手段と、を含むことを特徴とする決定リストの生成装置。

8.請求項5の装置により固定されたkに対して決定リストを求める手段と、該決定リストの末端から決定項を順次取り除くことによって得られる決定リストの系列を発生させる手段と、該系列要素の1つ1つに対して、その決定リストを用いて記述される観測データの記述量を計算する手段と、該記述量を記憶する手段と決定リストの系列を記憶する手段と、該決定リスト系列の中で記述量を最小にする決定リストを各kに対して最良の決定リス

トとして求める手段と、複数のkの値に対する最良の決定リストを記憶する手段と、複数のkの値に対する最良の決定リストの記述量を比較し、最小とする決定リストのを最終的に求める決定リストとして出力する手段と、を含むことを特徴とする決定リストの生成装置。

発明の詳細な説明

(産業上の利用分野)

この発明は複数の属性とクラスの組として与えられる、雑音を伴う複数の観測データから決定リストを発生させる方法に関する。

(従来の技術)

複数の属性とクラスの組として与えられる観測データから属性とクラスの構造的な関係を生成し、それを表現するための方法として決定リストによる分類規則生成の方法がある。決定リストの概念については、1987年発行の米国の雑誌「マシンラーンニング(Machine Learning)」の2巻の中の229-246頁掲載のR. Lリベスト(R. L. Rivest)による論文「ラーニングデシジョンリスト(Learnig decision lists)」に記載されており、2値データを扱う限りにおいては現在知られている分類規則の表現能力の高いものであることがわかっている。この論文の中では、雑音の伴わない観測データから、全ての観測データにつじつまを合わせるような決定リストを構成する方法が記載されている。

(発明が解決しようとする課題)

前記論文で示されていた決定リストの生成方法は、雑音の伴わない観測データから全ての観測データにつじつまを合わせるような決定リストを構成するための方法であった。しかし、実際に扱い観測データは一般に雑音や曖昧性などの不確実さを伴うので、全ての観測データを説明出来なくても、同じ情報源から発生するデータを出来るだけ正しく分類するような決定リストの方法が必要であるのだか、このような決定リストを発生させる手段は存在していなかった。

本発明の目的は雑音を伴う観測データから、未知データに対する予測誤差が最小になるような決定リストを自動的に発生させる方法を提供することにある。

(課題を解決するための手段)

本発明による決定リストの発生方法の1つは、1つの論理積を構成する文字の数を固定したもとで、決定リストの決定を規定する論理積として、観測データに対する情報利得を最大にするようなものを順次選択しながら付加するステップと、予め定めたストッピングルールによって決定の付加を終了したところで得られる決定リストを最終的に求める決定リストとして出力するステップとを含むことを特徴とする。

あるいは、上記発明に対して次のような変形を施すことも出来る。すなわち、上の方法によって一度決定リストを生成するステップと、次に該決定リストの末端から決定項を順次取り除くことによって得られる決定リストの系列の要素の1つ1つに対して、その決定リストを用いて記述される観測データの記述量を計算し、該決定リスト系列中で記述量を最小にする決定リストを最終的に求める決定リストとして出力するステップと、を含むことを特徴とする決定リストの発生方法である。

(上の2つの方法を「発明1」とする。)

また、本発明によるもう1つの決定リストの生成方法は、1つの論理積を構成する文字の最大数(kとする)を固定したもとで、決定リストの決定を規定する論理積として、観測データに対する情報利得を最大にするようなもの「順次選択しながら付加するステップと、予め定めたストッピングルールによって決定の付加を終了したところで得られる決定リストを複数のkの値に対して求めるステップと、各kに対して以上の生成過程を繰り返して生成される決定リストを用いて記述される観測データの記述量を計算して比較し、その中で記述量を最小にする決定リストを最終的に求める決定リストとして出力するステップと、を含むことを特徴とする。

あるいは、上記発明に対して次のような変形を施すことも出来る。すなわち、上の方法によって、各固定されたkに対して決定リストを求めるス

テップと、該決定リストの末端から決定項を順次取り除くことによって得られる決定リストの系列の要素の1つ1つに対して、その決定リストを用いて記述される観測データの記述量を計算し、該決定リスト系列の中で記述量を最小にする決定リストを各kに対して最良の決定リストとして求めるステップと、複数のkの値に対する最良の決定リストを求め、その中で記述量を最小とする決定リストを最終的に求める決定リストとして出力するステップと、を含むことを特徴とする決定リストの生成方法である。(以下、上の2つの方法を「発明2」とする)。

本発明による決定リストの生成装置の1つは、観測データを記憶する手段と、1つの論理積を構成する文字の数を固定したもとで、決定リストの決定を規定する論理積として観測データに対する情報利得を最大にするようなもの順次選択しながら付加し、予め定めたストッピングルールによって決定の付加を終了したところで得られる決定リストを最終的に求める決定リストとして出力する手段とを含むことを特徴とする。

あるいは、上記発明に対して次のような変形を施すことも出来る。すなわち、上の装置によって一度決定リストを生成する手段と、次に該決定リストの末端から決定項を順次取り除くことによって得られる決定リストの系列を発生される手段と、該系列要素の1つ1つに対して、その決定リストを用いて記述される観測データの記述量を計算する手段と、該記述量及び決定リストの系列を記憶する手段と、該決定リスト系列の中で記述量を最小にする決定リストを最終的に求める決定リストとして出力する手段と、を含むことを特徴とする決定リストの生成装置である。(上の2つの装置を「発明3」とする。)

また、本発明による決定リストの発生装置は、観測データを記憶する手段と、1つの論理積を構成する文字の最大数(kとする)を固定したもとで、決定リストの決定を規定する論理積として観測データに対する情報利得を最大にするようなものを順次選択しながら付加し、予め定めたストッピングルールによって決定の付加を終了したところで得られる決定リストを複数のkの値に対して求める手段と、各kに対して求められた決定リストを記憶する手段と、各kに対して以上の生成過程を繰り返して生成される決定リストを用いて記述される観測データの記述量を計算する手段と、該記述量を記憶する手段と、その中で記述量を最小にする決定リストを最終的に求める決定リストとして出力する手段と、を含むことを特徴とする

あるいは、上記発明に対して次のような変形を施すことも出来る。すなわち、上の装置により固定されたkに対して決定リストを求める手段と、該決定リストの末端から決定項を順次取り除くことによって得られる決定リストの系列を発生させる手段と、該系列要素の1つ1つに対して、その決定リストを用いて記述される観測データの記述量を計算する手段と、該記述量と決定リストの系列を記憶する手段と、該決定リストの中で記述量を最小にする決定リストを各kに対して最良の決定リストとして求める手段と、複数のkの値に対する最良の決定リストを記憶する手段と、複数のkの値に対する最良の決定リストの記述量を比較し、最小とする決定リストを最終的に求める決定リストを最終的に求める決定リストとして出力する手段と、を含むことを特徴とする決定リストの生成装置である。(上の2つの装置を「発明4」とする。)

(作用)

先ず、決定リストについて説明する。以下では簡単のため変数の値は全て{0, 1}であるとするが、一般には変数値は離散多値でも連続値であってもよい。今、{0, 1}に値をとる変数の数をnとし、kはn以下の正の整数であるとする。$L_n=\{x_1, \bar{x}_1, \cdots, x_n, \bar{x}_n\}(\bar{x}_i=1-x_i(i=1, \cdots, n))$とし、$L_n$の元をリテラルとよび、リテラルの論理積をタームとよぶ。k−DL(n)はT^n_kの元$t_j(j=1, \cdots, r)$と{0, 1}にとる値$v_j(j=1, \cdots, r)$の組のリストとして以下のように表されるものの集合を表す。ここに、T^n_kはL_nのうち高々異なるk個のリテラルで表されるターム全体の集合を示

し(但し、x_iと同時に$\bar{x}_i$を含まない)、最後の関数t_rはtrueを返す定数関数である。

$$(t_1, v_1), \cdots, (t_r, v_r) \quad (1)$$

1つの決定リストは任意の$x \in X_n$(X_nはn桁のブーリアンベクトルの全体の集合を表す)に対して$t_j(x)=1$となる最初のjに対するv_jの値を示す。(t_i, v_i)を第i決定(あるいは単に決定)と呼ぶ。例えば、次は3−DL(4)の元である。

$$(x_1\bar{x}_2, 1), (x_2\bar{x}_3\bar{x}_5, 0), (x_3x_4, 1), (\text{true}, 0) \quad (2)$$

この決定リストによるデータの分類の決定過程は第7図のように表される。

第7図で示したように、決定リストは"if−then−else　if−then−"型の概念分類規則の一般化となっている。n個の属性と1つのクラスで記述された複数の観測データから同じ情報源より発生するデータを出来るだけ正確に予測するような決定リストを構成するには、どのような方法あるいは装置によって決定リストを発生させたらよいかということが問題であり、この問題に答えるのが本発明である。但し、観測データには雑音や曖昧さなどの不確実性が含まれているものとする。先ず、この問題に答える1つの方法として、観測データに対する情報利得(エントロピー、誤分類率、Gini指標等の減少分で図られる)の高い決定から順に決定リストの決定として加えていくようにし、誤分類率に基づくストッピングルールを設けて、これが満足されたところで決定の追加を停止し、そこで得られる決定リストを出力するという方法、あるいはさらにそこで得られる決定リストの末端から決定項を順次取り除くことによって得られる決定リストの系列の要素の1つ1つに対して、その決定リストを用いて記述される観測データの記述量(全てを自己完結的に符号化するための符号長)を計算し、該決定リスト系列の中で記述量を最小にする決定リストを最終的に求める決定リストとする方法が考えられる。情報利得を最大にする決定から拾っていく理由は、識別能力が高い決定から先にデータを分類することにより、情報源から頻出するデータについて、より精度の高い分類・予測が行なえるからである。また、記述量を最小にする決定リストを選ぶことの理由は、観測データをモデルを用いて自己完結的な符号化を行う際に、全データをより圧縮しうるモデル(この場合は決定リスト)は、同じ情報源より発生する未知の観測データに対し、より正確な予測を行えるモデルであるということが漸近的に成立するからである。この理論的裏付けに関しては、特に統計モデルに適用する場合に関しては1986年発行の米国の雑誌アナルス オブ スタティスティクス(Annals of Statistics)の14巻の1080-1100頁掲載のJ. リサネン(J. Rissanen)による論文「ストキャスティック コンプレキシティー アンド モデリング(Stochastic complexty and modeling)」にMDL基準として述べられている。しかし、本発明では、分類規則としての決定リストといった論理型概念学習の最適モデル化にMDL基準を用いているところに新規性をおいている。以上が発明1の方法の原理であり、おなじ原理を装置の形で実現するのが発明3である。上述の情報利得を以下に正確に定義する。先ず、以下のような記号を定める。

[記号]　以下、対数の底は全て2とする。

- $k, n \in N$(自然数全体)は固定とする。
- $S \triangleq \{<i, x, c>$(学習データ): i(対象番号)$\in N$, x(属性値)$\in \{0,1\}^n$, c(クラス)$\in \{0,1\}\}$: 学習データの集合(属性値は属性変数$x_1, \cdots, x_n$に対するデータの値を示す。)
- $SA(t) = \{<i, x, c> \in S : t(x) = 1\}, t \in T^n_k$
  $SB(t) = \{<i, x, c> \in S : t(x) = 0\}, t \in T^n_k$
- $A(t) = \#SA(t)$(以下、#Wは集合Wの元の総数を表す)
  $B(t) = \#SB(t)$
- $A^+(t) = \#\{<i, x, c> \in SA(t) : c = 1\}, t \in T^n_k$
  $A^-(t) = \#\{<i, x, c> \in SA(t) : c = 0\}, t \in T^n_k$
- $B^+(t) = \#\{<i, x, c> \in SB(t) : c = 1\}, t \in T^n_k$
  $B^-(t) = \#\{<i, x, c> \in SB(t) : c = 0\}, t \in T^n_k$

ここに、$A(t) = A^+(t) + A^-(t)$, $B(t) = B^+(t) + B^-(t)$とする。

$$I(t) \triangleq \frac{A(t)}{A(t)+B(t)} I_A(t) + \frac{B(t)}{A(t)+B(t)} I_B(t) \quad (3)$$

$$\Delta I(t) \triangleq I_B(t'') - I(t) \quad (4)$$

を決定(t, v)による情報利得とよぶ。

(t"は決定リストにおいてtの直前に現れる決定のタームとする。)

ここで、$I_A(t), I_B(t)$の選び方としては次のようなものを考えることが出来る。

(i)エントロピー

$$I_A(t)\triangleq -\frac{A^+(t)}{A(t)}\log\frac{A^+(t)}{A(t)} -\frac{A^-(t)}{A(t)}\log\frac{A^-(t)}{A(t)}$$
$$I_B(t)\triangleq -\frac{B^+(t)}{B(t)}\log\frac{B^+(t)}{B(t)} -\frac{B^-(t)}{B(t)}\log\frac{B^-(t)}{B(t)} \quad (5)$$

(ii) Gini指標

$$\cdot I_A(t)\triangleq \frac{A^+(t)}{A(t)}\cdot\frac{A^-(t)}{A(t)}$$
$$\cdot I_B(t)\triangleq \frac{B^+(t)}{B(t)}\cdot\frac{B^-(t)}{B(t)} \quad (6)$$

(iii)誤分類率

$$\cdot I_A(t)\triangleq \frac{\min\{A^+(t), A^-(t)\}}{A(t)}$$
$$\cdot I_B(t)\triangleq \frac{\min\{B^+(t), B^-(t)\}}{B(t)} \quad (7)$$

情報利得$\Delta I(t)$の計算方法としては、上の他に、

$\Delta I(t)=-\{H(A^+(t)/A(t))+\delta\}\cdot 1/\log(A(t)+1)$

や

$\Delta I(t)=-\{H((A^+(t)+1)/(A(t)+2))+\delta\}\cdot 1/\log(A(t)+1)$

や

$\Delta I(t)=A(t)-A(t)\cdot H(A^+(t)/A(t))$

や

$\Delta I(t)=\{A(t)-A(t)\cdot H(A^+(t)/A(t))\}\cdot 1/\log(A(t)+1)$

などいろいろなものがとりうるものとする。但し、上に於て、δはある定数、$H(x)=-x\log x-(1-x)\log(1-x)$:エントロピー関数, $A(t)=A^+(t)+A^-(t)$とする。

次に、決定リストを用いた全観測データの記述量計算の例をあげる。データの記述量は、決定リスト自体の記述量と決定リストによって分類されるデータの中の例外の記述量との和あるいは何らかの補正を施して結合した量として与えられる。例えば、(2)の決定リストについては、$\#T^5_3=131$である(定数関数trueに対応するタームを0としてこれも数える)から、(2)の最初の決定のタームは1/131の確率で選ばれているので、これを自己完結的に符号化するためには、log(131)bitsの符号長が必要である。この場合、符号化方法としては、Huffman符号化を用いる。また、1つの決定に対しては、タームを真にするようなデータに1を割り付けるか0を割り付けるかも記述しなければならず、これには1bit必要であるから、結局(2)の最初の決定に関する記述量として、log(131)+1bits必要である。次の決定に関しては、T^5_3の最初の決定に用いられたタームを除く130個のタームの中から決定に必要なタームが選ばれるのであるから、同様にして、log(130)+1bitsの記述量が必要である。同様にして、各決定に対する記述量を求め、それらの総和を計算することにより決定リスト自体の記述量が計算できる。(2)に対しては全部で(log(131)+1)+(log(130)+1)+(log(129)+1)+(log(128)+1)bits必要である。また、例外データの記述量は、例えば、1つの決定に対して、その決定値が1であるとして、その決定におけるタームを真にするデータが7つであり、そのうちクラスが1であるものの数が5, 0であるものの数が2であるとして、対象番号の若い順にデータのクラスを記述すると、1101101であったとする。このとき、この系列を自己完結的に符号化するのに必要な符号長は、$\log(7+1)+\log({}_7C_2)$bitsまたは初めから例外は必ず$\lfloor(7+1)/2\rfloor$個以下であることが分かっていれば、$\log(\lfloor(7+1)/2\rfloor+1)+\log({}_7C_2)$bitsである。一般に、系列の長さをN, $b=b_1$またはb_2、ここに$b_1=N$, $b_2=\lfloor(N+1)/2\rfloor$とし、例外の数をhとすることにより、例外記述に必要な記述量は

$$\log(b+1)+\log({}_NC_h)\text{bits} \quad (8)$$

または、

$$L_N(h)+\log({}_NC_h)\text{bits} \quad (9)$$

で与えられる。ここに、$L_N(h)$は$\{0, 1, \cdots, N\}$に含まれる自然数を一意的に復号できるように符号化を行うときの符号長を表し、次に満たす。

$$L_N(0)=1$$
$$L_N(k)=1+\log k+\log\log k+\cdots C_N(k>1) \quad (10)$$

但し、上の和は正の項のみに対してとられるものであり、C_Nは

$$\sum_{k=0}^{M} 2^{-L_N(k)}=1$$

を満たす実数である。

記述量の計算は符号化の仕方に依存するので、上述の計算方法はあくまで例であり、これ以外にも、考えられる一意復号可能な符号化に対応して様々な記述長計算方法が用いられて良いものとする。

以上に与えた方法によって算出される決定リスト自身の記述量と例外データの記述量との和が決定リストを用いることによる全学習データの記述量である。あるいは決定リストの記述量と例外データの記述量のいずれかに補正を施した量を結合した利用を記述量として用いる場合もある。上で述べた情報利得と記述量の概念を用いて、所与の観測データに対する決定リストの最適化を行うことが出来る。例えば、属性数が6の第9図に示すような学習データを考える。これらの中に、たとえば、9と25のように属性値が同じでクラスが異なるといったデータの衝突が複数見られる。発明1によれば、先ず、kの値を固定したときの決定リストを、情報利得を最大にするタームtに対して、$A^+(t)\geqq A^-(t)$ならば(t, 1)を、そうでないならば(t, 0)を逐次的に付加して行く方法で構成し、予め定めたストッピングルールで停止して、そこで得られる決定リストを求める決定リストとして出力することにより、次のような決定リストが得られる(k=2, 3に対し、それぞれDL*(2), DL*(3)とかく)。但し、情報利得はエントロピーを用いるものとし、このときの決定付加の停止条件としてたとえば、前決定における決定値をv、情報利得を最大にするタームをtとして、次のi), ii), iii)の条件を採用するものとする。

i) $\min\{A^+(t), A^-(t)\}/A(t)>\alpha(=0.25)$ならば、
$B^+(t)\geqq B^-(t)$ならば(true, 1)を、そうでないならば(true, 0)を最終決定として付加して停止する。

ii) $\min\{B^+(t), B^-(t)\}/B(t)\leqq\beta(=0.29)$ならば、
$B^+(t)\geqq B^-(t)$ならば(true, 1)を、そうでないならば(true, 0)を最終決定として付加して停止する。

iii) まだ決定されていない観測データがもうなければ、
(true, 1−v)を決定として右に加えて出力し、停止する。

DL*(2)は次のとおり、
$(x_3\,x_5, 0)(x_2\,x_5, 0)(x_1\,x_4, 1)$(true, 1)
DL*(3)は次のとおり、
$(x_3, x_5, 0)(x_2\,x_5, 0)(x_1\,x_5\,x_6, 0)$(true, 1)
なお、上述のストッピグルールの決め方は、この方法にのみ限定はされない。

また、発明3によれば、観測データを記憶する手段と、情報利得を最大にする決定から順に付加して、予め定めているストッピグルールの条件が満たされたときに得られる決定リストを求める最終的な決定リストとして出力として出力する手段を具備している装置によって、k=2, 3のときにはそれぞれDL*(2), DL*(3)が発生させられる。

尚、以上の決定リストの発生方法並びに装置では、固定されたkに対してのみk−DL(n)が発生させることが出来る。一般にkの値が大きければ大きいほど表現能力が高くなり、多次元のデータ空間の分割の制度も細かくなるが、観測データには一般に雑音や曖昧さなどの不確実性をともなうが入っているので、kが必要以上に大きいと、統計的な揺らぎに過敏な決定リストを構成してしまうことになり、その場合、未知データに対してする分類予測誤差は返って大きくなる。従って、同じ情報源から発生する未知データに対する予測誤差を最小にするような最適なkの値が存在するはずであり、このようなkに対する決定リストを発生させる方法並びに装置が必要となる。上述の意味で最適な決定リストを発生させるためには、様々なkの値に対して、発明1の方法または発明は3の方法によって決定リストを発生させてから、それらを用いて記述される観測データの記述量を計算して比較し、記述量を最小にするような決定リストを最終的に求める決定リストとして求める方法並びに装置が考えられる。この理論的根拠も前出のJ. リサネンによる論文に示されているMDL基準の考え方に依るものである。以上が発明2の方法の原理であり、同

じ原理を装置の形で実現するのが発明4である。例えば、前出の例では、$\#T^6_2=73, \#T^6_3=233$であり、各決定における(決定される対象の数、例外の数)は、

DL*(2)で、

(6, 0), (3, 0), (7, 0), (19, 6)

DL*(3)で、

(6, 0), (3, 0), (2, 0), (24, 4)

であるから、記述量は、例外の記述量の計算方法としては$b=b_1$として(8)を用いることにすれば、次のように求められる。

DL*(2)で、

決定リストの記述量＝28.639bits

例外の記述量＝26.783bits

総記述量＝55.422bits

DL*(3)で、

決定リストの記述量＝35.419bits

例外の記述量＝24.411bits

総記述量＝59.831bits

従って、MDL基準の下ではDL*(2)がDL*(3)よりも適当なモデルであると言うことが出来る。発明2によると、k=2とk=3に対して発明1の方法で1度決定リストを発生させ、さらに各kに対する記述量を計算し、それらを比較して小さい記述量を与える決定リストとして、DL*(2)を発生させることが出来る。発明4によると、k=2とk=3に対して発明3の装置で1度決定リストを発生させる手段と、各kに対する記述量を計算する手段と、それら及び各kに対する決定リストを記憶する手段と、記述量を比較し、最小記述量をもつ決定リストを出力する手段を具備していれば、DL*(2)を発生させることが出来る。

(実施例)

次に、本発明について図面を参照して詳細に説明する。以下、記号は(作用)の項に従う。

第1図は発明1の実施例を説明するフローチャートである。startでは、対象番号と2値のn個の属性と2値のクラスからなる観測データの前集合を初期値とする集合Sと、固定されたn, kに対してT^n_kを初期値とする集合Tと、初期ターム$t''=0$(ターム0はT^n_kの中に含まれないが、全データを偽にするタームとして定める。また、このタームを用いた決定は出力時点では省略されるものとする)と、初期決定値v=0が与えられている。ステップ11で、ストッピングルールの条件として、最初から与えられている1以下の正の実数α, βに対して、

i) S＝空集合

ii) $\min\{A^+(t''), A^-(t'')\}/A(t'')>\alpha$

iii) $\min\{B^+(t''), B^-(t'')\}/B(t'')\leq\beta$

を設けて、i) ii) iii)の順にこれらの条件の1つでも当てはまるかどうかを調べる。これらの条件のうち一つでもあてはまるものがあれば、ステップ12に進み、最終決定として$B^+(t)\geq B^-(t)$ならば(true, 1)を、そうでないならば(true, 0)を付加して、それまで付加してきた決定と併せて出力して終了する。これらの条件がいずれも満たされていないならば、ステップ13に進み、(4)で定められている情報利得をタームtの関数とみなしたときにこれを最大にするタームtをTの中から決定する。(これをt^*とする。)情報利得最大のタームが複数ある場合は、その中でランダムにt^*を選ぶ。次に、ステップ14で、$A^+(t^*)>A^-(t^*)$ならばv=1とし、$A^+(t^*)<A^-(t^*)$ならばv=0とし、(t^*, v)を決定リストの決定として右に付け加える。次に、ステップ15で、タームtを真にするデータをSから除いた集合を改めてSとし、タームtをTから除いた集合を改めてTとし、t^*を改めてt''とし、ステップ11に戻る。

第2図は発明1のもう1つの実施例を説明するフローチャートである。ステップ21から25まではそれぞれ、第1図のステップ11から15までと同じ機能を果たし、ステップ25の出力として、第1図の出力と同じ決定リストが得られる。第2図では、ステップ25の出力としての決定リスト($DL_{max}(k)$とする)に対して、さらにステップ26で、$DL_{max}(k)$を用いて記述できる観測データの記述量を(作用)の項で述べたような方法で決定リスト自身の記述量と決定に対する例外データの記述量との和として計算し、これをL_0とする。ここで、記述量は、決定リスト

自身の記述量と決定に対する例外データの記述量との単純な和ではなく、重み付き和、すなわち、λ(決定リストの記述量)+(1−λ)(例外データの記述量)(0<λ<1)として計算はする場合あるいは決定リストの記述量と例外データの記述量のいずれかを補正して結合した量を用いる場合もあるとする。以下の記述量計算においても同様である。次に、ステップ27で、$DL_{max}(k)$の末端(最終)決定から刈り込みを行う。具体的には、$V(1)=DL_{max}(k)$として、j≧1に対し、V(j)の右から2つの決定を(t, v)(true, v')とするとき、これらを(true, v*)に置き換えたものをV(j+1)とする。但し、t'をtの直前の決定に用いられたタームとすると、v*は次のように与えられる。

$$v^*=\begin{cases}1 & (A^+(t)+B^+(t)\geqq A^-(t)+B^-(t)\text{の場合})\\ 0 & (\text{それ以外の場合})\end{cases}$$

次に、ステップ28で、V(j+1)を用いて記述される観測データの記述量を(作用)の項に示した方法で計算し、これをL_{j+1}とする。次に、ステップ29に進み、V(j+1)からさらに切り取る決定が残っているかどうかを判断し、残っていなければ、ステップ31に進み、記述量全てのL_jの値を比較して最小になるものに対するV(j)を出力し、終了する。ここで、もし、最小値が複数存在すれば、その最小値に対応するV(j)を複数出力し、終了する。また、ステップ29においてV(j+1)から刈り取る決定が残っていれば、ステップ30に進み、jをj+1としてステップ27に戻る。

第3図は発明2の実施例を説明するフローチャートである。startにおいては、対象番号と2値のn個の属性と2値のクラスとからなる観測データの全集合を初期値とする集合Sと、固定されたπ、と複数のkに対して(kは1つの決定に用いられる論理積を構成する文字の最大数であり、kのとりうる数の総数をMとする)、T^n_kを初期値とする集合T(k)と、初期ターム$t''=0$と、初期値決定値$v=0$が与えられている。

先ず、各kに対して、発明1の方法で決定リストを発生させる。$k_1<\cdots<k_M$を対象にするkとして、k_jに対応する決定リスト発生のステップを31とする。ステップ31は第1図の全体或は第2図の全体である。次に、発生した決定リストに対して、その決定リストを用いて記述されるデータの記述量を計算する。第2図で示された方法をステップ31にする場合はこのステップは省略される。k_iに対応する記述量計算のステップをステップ32で表す。次に、ステップ33各kの値に対して計算された記述量を比較して、最小値を求める。ここで、記述量は、決定リストの記述量と決定に対する例外データの記述量との単純な和ではなく、重み付き和、すなわち、λ(決定リストの記述量)+(1−λ)(例外データの記述量)(0<λ<1)として計算する場合、あるいは決定リストの記述量と例外データの記述量のいずれかを補正して結合した量を用いる場合もあるとする。次に、ステップ34で、これらの中で最小なkの値に対応する決定リストを出力し、終了する。第4図は発明3の装置を示すブロック図である。対象番号と2値のn個の属性と2個のクラスとからなる観測データの全集合を入力として、これを41の記憶装置で一旦記憶する。制御信号発生装置から発生する制御信号の指令によって記憶装置から観測データの一部が情報利得最大ターム決定回路42に送られる。最初に供給されるデータは観測データ全部である。情報利得最大ターム決定回路42は初期状態のパラメータとして、$T=T^n_k$(kは固定)、$t''=0$を有しており、

i) S=空集合

ii) $\min\{A^+(t''), A^-(t'')\}/A(t'')>\alpha$

iii) $\min\{B^+(t''), B^-(t'')\}/B(t'')\leqq\beta$

の条件をi) ii) iii)の順に調べ、1つでも当てはまるような場合は、信号P及びt''を42に従属する決定付加回路43に送る。また、上の条件がいずれも満たされない場合には、固定されたkの値に対して、Tの中のタームについて、記憶装置から供給されるデータに対して(4)で与えられる情報利得を計算し、これを最大にするようなタームt^*を決定し、状態を$t''=t^*$, $T=T-t^*$に変えて、t^*を決定付加回路43に送る。決定付加回路43は、情報利得最大ターム決定回路42の出力と記憶装置から供給されるデー

タ(42に供給されるデータと同じ)を入力とし、42からt*が送られてきた場合は$A^+(t^*), A^-(t^*)$を計算して、$A^+(t^*)>A^-(t^*)$ならば、v=1, $A^+(t^*)<A^-(t^*)$ならば、V=0とし、(t^*, v)を決定リストの決定として右に付け加え、次に記憶装置41がS−SA(t*)を42に供給するように指令する制御信号を発生させる信号を、制御信号発生装置44に送る。44は記憶装置がS−SA(t*)を42に供給するように指令する制御信号を発生させる信号を41に送る。44の指令を受けて新たに記憶装置41からデータが供給されると、42は、同じ動作を繰り返す。43に42から信号Pが送られてきた場合には、$B^+(t'') \geq B^-(t'')$ならば、v=1、そうでないならば、v=0として、(true, v)を最後の決定として決定リストの右に付け加えて、回路全体の出力としてこれまで決定を付加して得られた決定リストを出力し、回路全体の動作を終了する。

第5図は発明3の決定リストの発生装置のもう1つの実施例を示すブロック図である。回路51, 52, 53, 54はそれぞれ、第4図の回路41, 42, 43, 44と同じ入出力機能を果たする。しかし、回路53は、第4図の43の出力と同じ決定リストにくわえて、各決定に用いられたタームのそれぞれに対する$A^+(t), A^-(t)$の値も出力する。刈り込み及び記述量計算回路55は、回路53の出力を入力として、それを末端の決定から順に刈り込んで、それらを用いてデータを記述するときの記述量を併せて計算する。ここで、記述量は、決定リスト自身の記述量と決定に対する例外データの記述量との単純な和ではなく、重み付き和、すなわち、λ(決定リストの記述量)+(1−λ)(例外データの記述量)(0<λ<1)として計算する場合、あるいは決定リストの記述量と例外データの記述量のいずれかを補正して結合した値を用いる場合もあるとする。具体的には、右から2つの決定を(t, v)(true, v')とするとき、これらを(true, v*)に置き換えるといった操作を刈り取る決定が無くなるまで繰り返す。但し、v*は次のように与えられる。

$$V^* = \begin{cases} 1 & (A^+(t')+B^+(t') \geq A^-(t')+B^-(t')\text{の場合}) \\ 0 & (\text{そうでない場合}) \end{cases}$$

55はそれぞれ刈り込みによって得られた決定リストのそれぞれと、それぞれに対して計算された記述量を併せて、第2記憶装置56に送る。第2記憶装置は55から送られてきた入力を記憶する。また、55は刈り込まれたそれぞれの決定リストの記述量だけを記述量比較回路57にも送る。57は刈り込まれた決定リスト達の記述量を比較し、最小値を求め、第2記憶装置56から記述量を最小になる決定リストを出力させることを指示する制御信号を送ることを指令する信号を、第2制御信号発生回路58に送り込む。58は57の指令を受けて、第2記憶装置56から記述量を最小にする決定リストを出力させることを指示する制御信号を第2記憶装置56に送る。第2記憶装置56は58の指令を受けて、記述量を最小にする決定リストを出力する。

第6図は発明4の決定リストの発生装置の実施例を示すブロック図である。観測データを装置全体の入力として、該入力は先ず、DL*(k)発生回路61に送り込まれる。61は固定されたkに対して、発明3の装置と同じ回路によって決定リストを発生させ(固定されたkに対して発生された決定リストをDL*(k)とする)、これを第1制御信号発生装置62より発生する制御信号の指示に従って複数のkの値に対して実行する。61は各kに対するDL*(k)と各決定に用いられたタームのそれぞれに対する$A^+(t), A^-(t)$の値を順に記述量計算回路回路63に送り込み、各kに対するDL*(k)を記憶装置64に送り込む。63は61の出力を入力として、それぞれの記述量を計算し、その記述量を記述量比較回路65及び記憶装置64に送り込む。ここで、記述量は、決定リスト自身の記述量と決定に対する例外データの記述量との単純な和ではなく、重み付き和、すなわち、λ(決定リストの記述量)+(1−λ)(例外データの記述量)(0<λ<1)として計算する場合、あるいは決定リストの記述量と例外データの記述量のいずれかを補正して結合した量を用いる場合もあるとする。記憶装置64は61及び63の出力を入力として、これらを記憶する。65は63の出力を入力として、65は各kに対するDL*(k)の記述量を比較し、最小値を求

め、記憶装置64から記述量を最小にする決定リストを出力させることを指示する制御信号を送ることを指令する信号を、第2制御信号発生回路66に送り込む。66は65の指令を受けて、記憶装置64から記述量を最小にする決定リストを出力させることを指示する制御信号を記憶装置64に送る。記憶装置64は66の指令を受けて、記述量を最小にする決定リストを出力する。

(発明の効果)

発明1及び発明3によれば、固定されたkの値(決定を規定する論理積に現れる文字の最大数)に対して、観測データと同じ情報源から発生する未知データを高い制度で分類予測することが理論的に保証された決定リストが発生できる。発明2及び発明4によれば、さらに、kの値を動かして得られる広い決定リストの集合の中から、観測データと同じ情報源から発生する未知データを高い精度で分類予測することが理論的に保証され、また観測データをコンパクトに記述できる決定リストを選ぶことが出来る。

述量計算回路、64: 記憶装置、65: 記述量比較回路、66: 第2制御信号発生装置。

代理人 弁理士 内原 晋

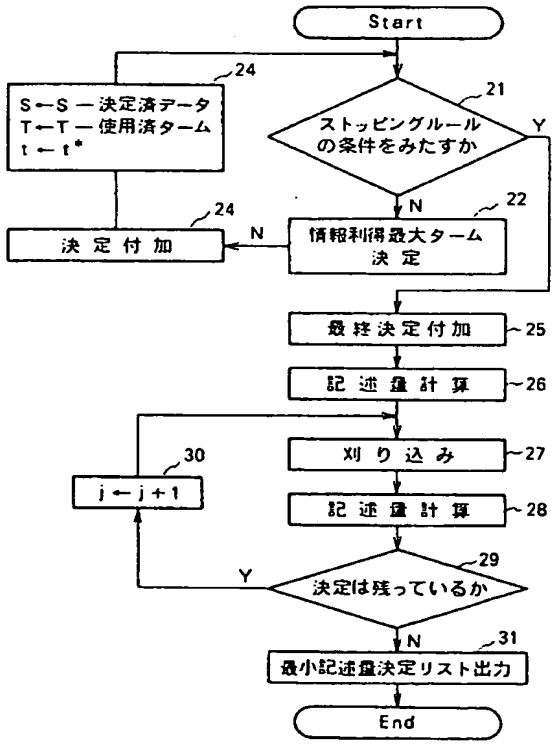
図面の簡単な説明

第1図は発明1の生成方法を示すフローチャート、第2図は発明1の決定リストの生成方法で、第1図に示した方法の変形版のフローチャート、第3図は発明2の決定リストの生成方法を示すフローチャート、第4図は発明3の決定リストの生成装置を示すブロック図、第5図は発明3の決定リストの生成装置で、第4図に示した方法の変形版のブロック図、第6図は発明4の決定リストの生成装置を示すブロック図、第7図、第8図、第9図は本発明の作用を説明するための図である。

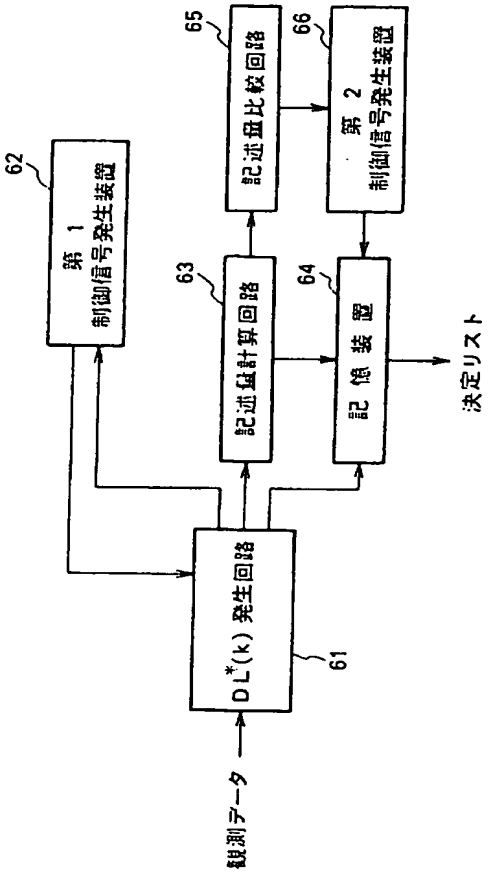
図において、

41: 記憶装置、42: 情報利得最大ターム決定回路、43: 決定付加回路、44: 制御信号発生装置、51: 第1記憶装置、52: 情報利得最大ターム決定回路、53: 決定付加回路、54: 第1制御信号発生装置、55: 刈り込み及び記述量計算回路、56: 第2記憶装置、57: 記述量比較回路、58: 第2制御信号発生装置、61: DL*(k)発生回路、62: 第1制御信号発生装置、63: 記

第 2 図



第 6 図



第 9 図 (a)

対象	x_1	x_2	x_3	x_4	x_5	x_6	クラス
1	0	1	1	0	1	1	1
2	1	1	1	0	0	0	1
3	0	1	0	1	1	0	1
4	0	1	1	0	0	1	1
5	1	1	1	0	1	1	1
6	1	0	0	0	0	1	0
7	0	1	1	1	1	0	1
8	1	0	1	0	1	1	0
9	0	1	0	1	1	0	0
10	1	1	1	0	1	1	1
11	0	1	1	0	0	1	1
12	0	1	1	1	0	0	0
13	1	0	0	0	0	0	0
14	0	1	1	1	1	1	1
15	0	1	1	0	1	1	1
16	1	1	1	0	0	1	0
17	0	1	1	0	1	1	1
18	0	1	0	1	0	1	0

第 9 図 (b)

対象	x_1	x_2	x_3	x_4	x_5	x_6	クラス
19	1	0	1	0	0	0	0
20	0	1	1	1	0	0	1
21	0	1	1	0	0	0	1
22	0	1	0	0	0	0	0
23	1	1	1	1	1	1	1
24	1	0	1	0	0	0	1
25	0	1	0	1	1	0	1
26	1	0	1	1	0	0	1
27	0	1	1	0	0	1	1
28	1	0	0	0	0	1	0
29	0	1	1	0	0	0	1
30	0	1	0	1	1	0	1
31	0	0	0	0	0	1	0
32	1	1	1	1	1	0	0
33	1	0	1	0	1	1	0
34	1	1	0	1	1	0	0
35	1	0	1	0	1	1	0